Advances in Applied Mathematics 应用数学进展, 2026, 15(2), 247-256
Published Online February 2026 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2026.152066

Hans汉斯

基于PiE正则化与自步学习的稀疏逻辑回归特征选择方法

马娟玲

广东工业大学数学与统计学院，广东 广州

收稿日期：2026年1月5日；录用日期：2026年1月29日；发布日期：2026年2月5日

摘　要

针对高维小样本、高噪声基因表达数据的特征选择与分类问题，本文提出一种融合分段指数(PiE)正则化与自步学习(SPL)机制的稀疏逻辑回归模型(PiE-SPLR)。PiE正则化能逼近l_0范数，具有强稀疏选择能力；SPL机制可逐步筛选低噪声样本，增强模型鲁棒性。通过交替方向优化与近端梯度法高效求解，模型在Colon、Leukemia等四个基因数据集上取得了最优分类性能，同时所选特征数最少。该方法为基因标志物挖掘与高维数据分类提供了有效工具。

关键词

PiE正则化，自步学习，稀疏逻辑回归，特征选择，基因表达数据

Sparse Logistic Regression Feature Selection Method Based on PiE Regularization and Self-Paced Learning

Juanling Ma

School of Mathematics and Statistics, Guangdong University of Technology, Guangzhou Guangdong

Received: January 5, 2026; accepted: January 29, 2026; published: February 5, 2026

Abstract

To address the feature selection and classification challenges in high-dimensional, small-sample, and high-noise gene expression data, this paper proposes a sparse logistic regression model (PIE-SPLR) that integrates Piecewise Exponential (PiE) regularization with Self-Paced Learning (SPL).

文章引用: 马娟玲. 基于 PiE 正则化与自步学习的稀疏逻辑回归特征选择方法[J]. 应用数学进展, 2026, 15(2): 247-256.
DOI: 10.12677/aam.2026.152066

The PiE regularization approximates the l_0 norm, providing strong sparse selection capability, while the SPL mechanism progressively filters low-noise samples to enhance model robustness. Through efficient alternating direction optimization and proximal gradient methods, the model achieves optimal classification performance on four gene datasets (Colon, Leukemia, etc.), while selecting the fewest features. This method provides an effective tool for biomarker discovery and high-dimensional data classification.

Keywords

PiE Regularization, Self-Paced Learning, Sparse Logistic Regression, Feature Selection, Gene Expression Data

1. 引言

现代高通量生物医学仪器的广泛应用极大地加快了生命科学领域的数据生成速度。例如，美国国家生物技术信息中心的基因表达综合数据库已收集了超过三百万个样本。基因组研究的核心任务之一，即是预测表型并发现少量关键生物标志物[1] [2]。

然而，运用机器学习方法分析基因组数据(例如基因表达数据)时，主要面临两大挑战：1) “大 p 小 n”问题。基因表达数据集通常包含大量基因(p)和少量样本(n)，其中仅有极少数基因与目标疾病真正相关。大量无关特征会引入噪声、导致模型过拟合，并严重损害分类器的泛化性能与可解释性[3] [4]。2) 高噪声问题。在生物数据生成过程中，无论是样本制备、实验操作还是批次效应，都会引入显著噪声，使模型训练不稳定、预测结果不可靠[5]。

特征选择是应对上述问题的关键。其中，正则化方法通过约束模型参数，能够同时实现连续收缩与自动特征选择，是处理高维小样本数据的有效手段[6]。经典的 l_0 正则化(Lasso) [7]及其扩展(如 SCAD [8]、Elastic Net [9])虽被广泛使用，但所产生的解往往不够稀疏，且在噪声较大时估计偏差明显[10]。近年来，非凸正则化因其更强的稀疏诱导能力受到关注。Xu 等人提出的 $l_{1/2}$ 正则化[11]在理论性质和计算效率之间取得了较好平衡，并具有 Oracle 性质[11] [12]。然而，直接使用 $l_{1/2}$ 范数仍是对 l_0 范数的一种近似，其稀疏能力与理论最优的 l_0 约束之间尚有差距。文章[13]总结并列举出了许多 l_0 范数的近似函数。在许多近似函数中，分段指数(PiE)函数因为可以比 l_1 范数产生更大的稀疏性而引起了极大关注。

另一方面，为提升模型在噪声数据下的鲁棒性，自步学习(Self-Paced Learning, SPL)机制被引入[14]。它通过逐步从易到难选择样本参与训练，有效降低了噪声样本对模型更新的负面影响，已在多种学习任务中展现出优势[15] [16]。但如何将 SPL 与更具稀疏潜力的正则化方法相结合，仍有待深入探索。

针对上述问题，本文提出一种基于 PiE 函数的自步学习稀疏逻辑回归模型。该模型的创新性主要体现在以下两方面：

1) 引入新型稀疏正则化函数：我们采用 PiE 函数作为 l_0 范数的近似。理论分析表明，该函数比 $l_{1/2}$ 等经典非凸正则子具有更贴近 l_0 的几何形态与数学性质，能在保证优化可行性的同时，诱导出更稀疏、更稳定的特征选择结果。

2) 构建抗噪声的联合学习框架：我们将 PiE 正则化与自步学习机制深度融合，构建了一个统一的目

标函数与优化算法。该框架能够同步实现高维特征选择与噪声样本自适应加权，从而在“大 p 小 n”与高噪声并存的数据环境下，获得更具可解释性的生物标志物集合与更鲁棒的表型分类模型。

为验证所提方法的有效性，我们在多个公共基因表达数据集上进行了实验，并与 Lasso、SCAD、$l_{1/2}$ 等方法进行比较。结果表明，新模型在保持较高分类精度的前提下，能够选择更少且更具生物相关性的基因，并显著提升在高噪声场景下的预测稳定性。

2. 模型介绍

针对高维小样本基因表达数据中的特征稀疏与噪声干扰问题，本章提出一种基于 PiE 正则化的自适应步学习逻辑回归模型(PiE-SPLR)。该模型从损失函数、稀疏约束与抗噪学习三个层面进行系统整合：以逻辑回归为基础分类器；引入 PiE 函数近似 l_0 范数，增强特征选择的稀疏性；并融合自适应步学习机制，通过动态样本加权提升模型在高噪声下的鲁棒性。下文将对各部分进行详细介绍。

2.1. 逻辑回归损失函数

逻辑回归是一种广泛应用于二分类问题的统计学习方法。对于给定的训练数据集 $\mathcal{D} \coloneqq \{(x_i, y_i) : i \in [n] \subseteq \mathbb{R}^p \times \{0,1\}\}$，其中 $[n] \coloneqq \{1,2,\cdots,n\}$。逻辑回归通过 Sigmoid 函数将线性组合 $x_i^\top \beta$ 映射为样本属于正例($y_i = 1$)的预测概率：

$$P(y_i = 1 \mid x_i; \beta) = \sigma(x_i^\top \beta) = \frac{1}{1 + \mathrm{e}^{-x_i^\top \beta}},$$

其中 $\beta \in \mathbb{R}^p$ 为待估计的模型系数向量。

为了拟合模型参数，逻辑回归采用极大似然估计，其对应的损失函数(即负对数似然)为：

$$\ell(\beta) = -\sum_{i=1}^{n} \ell_i(\beta) = -\sum_{i=1}^{n} \left[y_i \log \sigma(x_i^\top \beta) + (1 - y_i) \log\left(1 - \sigma(x_i^\top \beta)\right) \right]. \tag{2.1}$$

式(2.1)衡量模型预测概率与真实标签之间的差异，通过最小化式(2.1)可获得系数 β 的估计值，从而构建分类决策边界。

在本研究中，逻辑回归损失函数将作为模型的基础预测模块，后续通过引入正则化项与样本加权机制，增强其在高维特征选择与噪声数据拟合中的性能。

2.2. 正则化函数

本文采用 PiE 函数作为正则项，其表达式为：

$$P(\beta) = \sum_{i=1}^{p} p(\beta_i) \coloneqq \sum_{i=1}^{p} \left(1 - \mathrm{e}^{-\frac{|\beta_i|}{\sigma}}\right), \ \forall \beta \in \mathbb{R}^p. \tag{2.2}$$

如图 1 所示，不同 σ 取值下 PiE 函数对 l_0 范数的逼近效果存在显著差异。当 σ 值较小时(如 $\sigma = 0.01$)，函数曲线更为陡峭，能更紧密地逼近 l_0 范数的理想阶跃特性；随着 σ 值增大(如 $\sigma = 1.0$)，函数曲线逐渐平滑，稀疏诱导能力相应减弱。由此可见，参数 σ 是控制稀疏性与优化可行性的关键平衡因子。

2.3. SPL

面对高噪声生物数据对模型训练的干扰，SPL 提供了一种有效的样本重加权策略。与传统优化方法对所有样本平等对待不同，SPL 模拟人类由易到难的学习认知过程，通过在训练过程中动态调整样本权重，使模型优先学习高置信度的“简单”样本，再逐步纳入更具挑战性的样本，从而提升模型在噪声环境下的鲁棒性与泛化能力[14]。

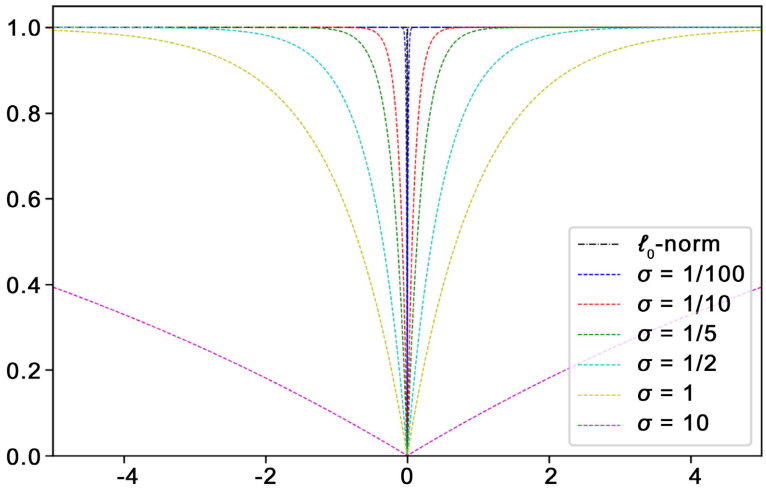

Figure 1. The approximation relationship between the PiE function and the l_0 norm under different values of σ

图 1. 不同 σ 取值下 PiE 函数对 l_0 范数的逼近关系

经典的 SPL 框架通常表述为以下优化问题：

$$\min_{\beta,v} E(\beta,v;\tau)=\sum_{i=1}^{n}\left[v_i\ell_i(\beta)+f(v_i;\tau)\right].$$

其中 $v=(v_1,v_2,\cdots,v_n)^{\top}\in[0,1]^n$ 为样本权重向量，$\ell(\cdot)$ 为损失函数，$\tau>0$ 为年龄参数，控制学习进程的“步速”。$f(v_i;\tau)$ 为自步惩罚项，用于调控样本权重的分配，其常见形式为线性硬惩罚 $f(v_i;\tau)=-\tau v_i$。

2.4. PiE-SPLR 模型

基于前文所述，本文将逻辑回归损失函数、PiE 正则化函数与自适应步学习机制相结合，构建 PiE 正则化自适应步学习逻辑回归模型(PiE-SPLR)。该模型的目标函数如下：

$$\min_{\beta,v}\mathcal{L}(\beta,v)=\sum_{i=1}^{n}\left[v_i\ell_i(\beta)+f(v_i;\tau)\right]+\lambda P(\beta). \tag{2.3}$$

其中，第一项为加权逻辑回归损失函数，第二项 $f(v_i;\tau)=-\tau v_i$ 为自适应步惩罚项，$v_i\in[0,1]$为样本权重，$\lambda>0$ 为正则化参数。

3. 优化算法

3.1. 交替方向优化算法

为求解式(2.3)定义的 PiE-SPLR 模型，本文采用交替方向优化(Alternating Direction Optimization)策略。该算法将原联合优化问题分解为两个相对简单的子问题：固定样本权重 v 更新模型参数 β，以及固定模型参数 β 更新样本权重 v。两个子问题交替迭代直至收敛。

3.2. 模型参数 β 的更新

固定样本权重 v，关于 β 的优化问题简化为：

$$\min_{\beta}\mathcal{L}_{\beta}(\beta)=\sum_{i=1}^{n}v_i\ell_i(\beta)+\lambda P(\beta). \tag{3.1}$$

式(3.1)包含可微的逻辑回归损失项与非凸、非光滑的 PiE 正则项。本文采用近端梯度下降(Proximal Gradient Descent)算法进行求解。具体地，在每次迭代中，首先计算损失函数的梯度步，然后应用 PiE 函数的近端算子(Proximal Operator)。

设第 t 次迭代的参数为 $\beta^{(t)}$，学习率为 $\eta>0$，则更新步骤为：

1) 梯度计算：

$$g^{(t)}=\nabla_{\beta}\left(\sum_{i=1}^{n}v_i\ell_i(\beta)\right)\Big|_{\beta=\beta^{(t)}}=\sum_{i=1}^{n}v_i^{(k)}\left(\ell_i\left(\beta^{(k)}\right)-y_i\right)x_i. \tag{3.2}$$

2) 梯度步：

$$\tilde{\beta}^{(t)}=\beta^{(t)}-\eta g^{(t)}. \tag{3.3}$$

3) 近端映射：

$$\beta^{(t+1)}=\mathrm{prox}_{\eta\lambda P}\left(\tilde{\beta}^{(t)}\right). \tag{3.4}$$

其中，$\mathrm{prox}_{\eta\lambda P}(\cdot)$ 是 PiE 函数的近端算子，定义为：

$$\mathrm{prox}_{\eta\lambda P}(z)=\arg\min_{\beta}\left\{\frac{1}{2\eta}\|\beta-z\|^2+\lambda P(\beta)\right\}$$

PiE 函数的近端算子已在文章[17]中推导出来，假设 η，σ，$\lambda>0$，其显式表达式为：

1) 如果 $\eta\lambda\le\sigma^2$，则对于任意的 $z\in\mathbb{R}$，有

$$\mathrm{prox}_{\eta\lambda P}(z)=\begin{cases}\{0\}, & \text{若 } |z|\le\dfrac{\eta\lambda}{\sigma},\\ \left\{sign(z)\sigma W_o\left(-\dfrac{\eta\lambda}{\sigma^2}\right)\mathrm{e}^{-\frac{|z|}{\sigma}}+z\right\}, & \text{其它,}\end{cases}$$

2) 如果 $\eta\lambda>\sigma^2$，则对于任意的 $z\in\mathbb{R}$，有

$$\mathrm{prox}_{\eta\lambda P}(z)=\begin{cases}\{0\}, & \text{若 } |z|<\bar{z}_{\eta\lambda,\sigma},\\ \left\{0, sign(z)\sigma W_o\left(-\dfrac{\eta\lambda}{\sigma^2}\right)\mathrm{e}^{-\frac{|z|}{\sigma}}+z\right\}, & \text{若 } |z|=\bar{z}_{\eta\lambda,\sigma},\\ \left\{sign(z)\sigma W_o\left(-\dfrac{\eta\lambda}{\sigma^2}\right)\mathrm{e}^{-\frac{|z|}{\sigma}}+z\right\}, & \text{其它,}\end{cases}$$

其中，阈值 $\bar{z}_{\eta\lambda,\sigma}=z^{\star}+\dfrac{\eta\lambda}{\sigma}\mathrm{e}^{-\frac{z^{\star}}{\sigma}}$，$z^{\star}\in\left(0,\sqrt{2\eta\lambda}\right)$ 是方程 $\dfrac{1}{2}+\eta\lambda\dfrac{\left(\frac{z}{\sigma}+1\right)\mathrm{e}^{-\frac{z}{\sigma}}-1}{z^2}=0$ 的唯一解。函数 W_0 是满足 $W(z)\ge-1$ 的 Lambert W 函数 $W(z)$ 的分支[18]。

3.3. 样本权重 v 的更新

固定模型参数 β，关于样本权重 v 的优化问题可分解为 n 个独立的子问题：

$$\min_{v_i\in[0,1]}\left\{v_i\cdot\ell_i(\beta)+f(v_i;\tau)\right\},\ i=1,\cdots,n. \tag{3.5}$$

其中 $\ell_i(\beta)$ 表示第 i 个样本在当前模型下的损失值。采用硬惩罚函数 $f(v_i;\tau)=-\tau v_i$，则式(3.5)可重写为：

$$\min_{v_i\in[0,1]}\left\{v_i\left(\ell_i(\beta)-\tau\right)\right\}. \tag{3.6}$$

式(3.6)是一个关于 v_i 的线性规划问题。其最优解可通过分析目标函数的性质直接得到，从而获得样本权重 v_i 的闭式更新公式：

$$v_i^{(t+1)}=\begin{cases}1, & \text{若 } \ell_i\left(\beta^{(t+1)}\right)\le\tau^{(t)},\\ 0, & \text{若 } \ell_i\left(\beta^{(t+1)}\right)>\tau^{(t)},\end{cases}\quad i=1,\cdots,n. \tag{3.7}$$

其中 $\ell_i\left(\beta^{(t+1)}\right)$ 是基于最新模型参数第 i 个样本损失，$\tau^{(t)}$ 是当前迭代的年龄参数。式(3.7)定义了一个硬阈值筛选规则。仅当样本的损失小于当前年龄参数 $\tau^{(t)}$ 时，该样本才会被纳入训练($v_i=1$)；否则被暂时排除($v_i=0$)。年龄参数 τ 在每轮迭代后按线性增长策略更新：

$$\tau^{(t+1)}=\rho\tau^{(t)}. \tag{3.8}$$

其中 $\rho>1$ 为增长因子。随着 τ 的增大，满足 $\ell_i(\beta)<\tau$ 条件的样本逐渐增多，模型逐步纳入更多“困难”样本。

3.4. 完整算法流程

基于上述更新规则，完整的 PiE-SPLR 优化算法如表 1 所示。

Table 1. PiE-SPLR optimization algorithm
表 1. PiE-SPLR 优化算法

步骤	操作	说明
输入	数据 $\{(x_i,y_i)\}_{i=1}^n$；$\lambda,\tau^{(0)},\rho,\eta,T,\epsilon$	λ：正则化参数；$\rho>1$：增长因子
1	初始化：$t=0$，$\beta^{(0)}$，$v^{(0)}$	设置初始值
2	**while** $t<T$ **and** $\|\Delta\beta\|>\epsilon$ **do**	主循环
3	计算梯度 $g^{(t)}$ (3.2)和梯度步 $\tilde{\beta}^{(t)}$ (3.3)	
4	更新 β：$\beta^{(t+1)}=\mathrm{prox}_{\eta\lambda P}\left(\tilde{\beta}^{(t)}\right)$ (3.4)	近端梯度步
5	更新 v (3.7)	硬阈值筛选
6	更新 τ (3.8)	自适应阈值
7	$t\leftarrow t+1$	迭代计数
8	**end while**	
输出	$\hat{\beta}\leftarrow\beta^{(t)}$，$\hat{v}\leftarrow v^{(t)}$	最终模型

4. 实验结果与分析

4.1. 实验设置与对比方法

本章在四个公开的基因表达数据集上验证所提 PiE-SPLR 模型的有效性。选用的数据集包括：Colon [19]、Leukemia [20]、DLBCL [21]和 Prostate [22]，表 2 列出了这些数据集的更多信息。这些数据集均代表典型的高维小样本基因数据，且广泛应用于特征选择方法评估。

Table 2. Statistical information of experimental datasets
表 2. 实验数据集统计信息

数据集	样本数	特征数	类别分布	
			正常/阴性	肿瘤/阳性
Colon	62	2000	22	40
Leukemia	72	3571	25 (AML)	47 (ALL)
DLBCL	77	7129	19 (FL)	58 (DLBCL)
Prostate	102	12,600	50	52

为全面比较性能，选取了六种代表性方法作为基准：Lasso (l_1 正则化逻辑回归)、$l_{1/2}$ ($l_{1/2}$ 正则化逻辑回归)、SCAD-l_2 (SCAD 正则化逻辑回归)、l_1-Net (l_1 网络正则化逻辑回归)、Inter-Net (交互网络正则化逻辑回归)、SLNL [23] ($l_{1/2}$ - NL 正则化逻辑回归)以及 $l_{1/2}$-Net ($l_{1/2}$ 网络正则化逻辑回归)。所有方法均采用相同的 5 折交叉验证流程进行参数调优与性能评估。评价指标主要包括测试集分类训练准确率(Training Acc)、分类测试准确率(Testing Acc)、以及模型选择的特征数量(Genes)。PiE-SPLR 的主要参数通过网格搜索确定，其中正则化参数 λ 从区间 $\{1:0.1:2\}$ 中选取，初始年龄参数 $\tau^{(0)}$ 设为 0.1，增长因子 ρ 从区间 $\{1:0.05:1.2\}$ 中选取，学习率 $\eta \in \{0.02, 0.03, 0.04\}$，收敛容差 $\epsilon = 10^{-5}$，最大迭代次数 $T = 800$。

4.2. 分类性能与特征选择结果

表 3 和表 4 分别展示了各对比方法在训练集和测试集上的分类准确率。从训练结果可见，本文提出的 PiE-SPLR 在所有四个数据集上均达到最高准确率，其中在 DLBCL 数据集上更是取得了接近完美的 99.99%准确率。特别值得注意的是，PiE-SPLR 在训练性能上相较原 SLNL 方法有明显提升，例如在 Colon 数据集上从 93.81%提升至 94.02%，在 Leukemia 数据集上从 98.75%提升至 99.95%。

更为重要的是测试集上的表现(表 4)，这直接反映了模型的泛化能力。PiE-SPLR 在所有数据集上均取得了最优测试准确率，平均达到 96.58%，较次优方法 SLNL (平均 92.65%)提升 3.93 个百分点。其中在 DLBCL 数据集上的提升尤为显著，从 91.56%跃升至 98.47%。这一结果表明，PiE-SPLR 不仅在训练过程中能更好地拟合数据，更重要的是具备了更强的泛化能力，避免了过拟合现象。

Table 3. Comparison of Training Acc among different methods (%)
表 3. 各方法 Training Acc 比较(%)

方法	Colon	Leukemia	DLBCL	Prostate
Lasso	88.61	97.00	94.95	93.29
$l_{1/2}$	89.52	93.87	92.69	92.29
SCAD-l_2	90.47	94.06	94.80	92.04
l_1-Net	88.99	98.45	94.13	93.23
Inter-Net	87.80	98.14	95.32	93.54
$l_{1/2}$-Net	91.46	96.36	96.48	94.61
SLNL	93.81	98.75	97.55	95.72
PiE-SPLR	**94.02**	**99.95**	**99.99**	**97.78**

Table 4. Comparison of Training Acc among different methods (%)
表 4. 各方法 Testing Acc 比较(%)

方法	Colon	Leukemia	DLBCL	Prostate
Lasso	80.23	93.41	84.89	88.81
$l_{1/2}$	85.16	93.84	88.81	89.73
SCAD- l_2	82.29	95.36	88.93	89.84
l_1 -Net	86.40	95.99	87.25	88.48
Inter-Net	85.58	94.13	87.51	90.25
$l_{1/2}$ -Net	86.84	95.54	89.48	90.51
SLNL	87.46	97.85	91.56	93.73
PiE-SPLR	**94.78**	**99.14**	**98.47**	**93.91**

Table 5. Comparison of selected Genes among different methods
表 5. 各方法 Genes 比较

方法	Colon	Leukemia	DLBCL	Prostate
Lasso	8.3	6.6	23.1	24.7
$l_{1/2}$	8.0	**3.9**	14.8	14.5
SCAD- l_2	18.7	15.3	33.5	29.9
l_1 -Net	22.8	14.7	40.2	47.4
Inter-Net	26.1	20.2	37.4	51.3
$l_{1/2}$ -Net	21.2	8.8	30.1	17.7
SLNL	17.7	9.2	24.7	21.3
PiE-SPLR	**7.69**	6.76	**12.27**	**11.3**

在特征选择方面，表 5 呈现了各方法选择的特征数量。PiE-SPLR 展现出更好的稀疏性，在所有数据集上均选择了最少的特征数，平均仅为 9.51 个。这一结果远低于原 SLNL 方法的平均 18.23 个特征，更显著少于其他对比方法。值得注意的是，尽管 PiE-SPLR 选择了更少的特征，但其分类性能却得到了全面提升，这验证了 PiE 正则化在逼近 l_0 范数方面的优势：它能够更精确地识别并保留真正关键的生物标志物，同时更有效地剔除冗余和不相关特征。

噪声环境下的进一步验证

为进一步验证 SPL 机制对噪声的鲁棒性，我们在 Colon 数据集中引入 20%标签噪声进行补充实验。如表 6 所示，PiE-SPLR 在噪声下仍保持最优性能，其关键基因保留率(KGRR)高达 85%，远高于 PiE-LR (62%)和 Lasso (45%)。这证实 SPL 机制通过动态样本加权，能有效过滤噪声样本，提升特征选择的稳定性。

4.3. 结果讨论与总结

综合以上实验结果，PiE-SPLR 的优越性能可归结于其创新性的双重选择机制。PiE 正则化通过对 l_0 范数的紧致逼近，实现了比传统 $l_{1/2}$ 正则化更精确的稀疏控制，避免了因过度稀疏而丢失重要特征的问题。同时，在标签噪声环境下仍能保持高准确率与高稳定性，这得益于 SPL 机制对噪声样本的自适应过滤能

力，为高噪声生物医学数据的分析提供了可靠工具。自适应步学习通过动态调整样本权重，由“简单”到“复杂”样本的训练过程，有效过滤了噪声样本的干扰。这两者的协同作用使模型在特征选择与噪声鲁棒性之间取得了良好平衡。

Table 6. Comparison of feature selection stability under noisy environment
表 6. 噪声环境下特征选择稳定性对比

方法	测试准确率(%)	选中基因数	KGRR (%)
PiE-SPLR	**88.56**	**8.2**	**85.0**
PiE-LR	79.24	15.6	62.0
Lasso	72.31	22.8	45.0

本章实验全面验证了 PiE-SPLR 在高维小样本基因数据分类任务中的有效性。该方法在分类准确率、特征选择质量和噪声鲁棒性三个关键指标上均表现出显著优势，为基因表达数据分析提供了一种新的有效工具。在后续工作中，我们将进一步探索该方法在其他组学数据整合分析中的应用潜力，并考虑将其扩展到多类别分类和生存分析等更复杂的生物医学问题中。

参考文献

[1] Song, X., Liu, M.T., Liu, Q. and Niu, B. (2021) Hydrological Cycling Optimization-Based Multiobjective Feature-Selection Method for Customer Segmentation. *International Journal of Intelligent Systems*, **36**, 2347-2366. https://doi.org/10.1002/int.22381

[2] Li, C.-N., Shao, Y.-H., Zhao, D., Guo, Y.-R. and Hua, X.-Y. (2020) Feature Selection for High-Dimensional Regression via Sparse LSSVR Based on LP-Norm. *International Journal of Intelligent Systems*, **36**, 1108-1130.

[3] Frankell, A.M., Jammula, S., Contino, G., Killcoyne, S.S. and Fitzgerald, R.C. (2018) The Landscape of Selection in 551 Esophageal Adenocarcinomas Defines Genomic Biomarkers for the Clinic. Nature Genetics.

[4] Huang, H.H., Peng, X.D. and Liang, Y. (2021) Splsn: An Efficient Tool for Survival Analysis and Biomarker Selection. *International Journal of Intelligent Systems*, **36**, 5845-5865.

[5] Fei, T. and Yu, T. (2020) Scbatch: Batch-Effect Correction of RNA-Seq Data through Sample Distance Matrix Adjustment. *Bioinformatics*, **36**, 3115-3123. https://doi.org/10.1093/bioinformatics/btaa097

[6] Armanfard, N., Reilly, J.P. and Komeili, M. (2016) Local Feature Selection for Data Classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **38**, 1217-1227. https://doi.org/10.1109/tpami.2015.2478471

[7] Tibshirani, R. (1996) Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society Series B*: *Statistical Methodology*, **58**, 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

[8] Zhang, X., Wu, Y., Wang, L. and Li, R. (2015) Variable Selection for Support Vector Machines in Moderately High Dimensions. *Journal of the Royal Statistical Society Series B*: *Statistical Methodology*, **78**, 53-76. https://doi.org/10.1111/rssb.12100

[9] Wang, L., Zhu, J. and Zou, H. (2006) The Doubly Regularized Support Vector Machine. *Statistica Sinica*, 589-615.

[10] Meinshausen, N. and Yu, B. (2009) Lasso-Type Recovery of Sparse Representations for High-Dimensional Data. *The Annals of Statistics*, **37**, 246-270. https://doi.org/10.1214/07-aos582

[11] Xu, Z., Zhang, H., Wang, Y., Chang, X. and Liang, Y. (2010) $l_{1/2}$ Regularization. *Science China*: *Information Sciences*, **53**, 1159-1169.

[12] Zeng, J., Lin, S., Wang, Y., Xu, and Zongben. (2014) $l_{1/2}$ Regularization: Convergence of Iterative Half Thresholding Algorithm. *IEEE Transactions on Signal Processing*, **62**, 2317-2329.

[13] Xu, F., Duan, J. and Liu, W. (2023) Comparative Study of Non-Convex Penalties and Related Algorithms in Com-Pressed Sensing. *Digital Signal Processing*, **135**, 103937.

[14] Kumar, M.P., Packer, B. and Koller, D. (2010) Self-Paced Learning for Latent Variable Models. Curran Associates Inc.

[15] Li, C., Wei, F., Yan, J., Zhang, X., Liu, Q. and Zha, H. (2018) A Self-Paced Regularization Framework for Multilabel Learning. *IEEE Transactions on Neural Networks and Learning Systems*, **29**, 2660-2666.

https://doi.org/10.1109/tnnls.2017.2697767

[16] Huang, H. and Liang, Y. (2019) An Integrative Analysis System of Gene Expression Using Self-Paced Learning and Scad-Net. *Expert Systems with Applications*, **135**, 102-112. https://doi.org/10.1016/j.eswa.2019.06.016

[17] Liu, Y., Zhou, Y. and Lin, R. (2024) The Proximal Operator of the Piece-Wise Exponential Function. *IEEE Signal Processing Letters*, **31**, 894-898. https://doi.org/10.1109/lsp.2024.3370493

[18] Mező, I. (2022) The Lambert W Function: Its Generalizations and Applications. Chapman and Hall/CRC. https://doi.org/10.1201/9781003168102

[19] Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D., *et al*. (1999) Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays. *Proceedings of the National Academy of Sciences*, **96**, 6745-6750. https://doi.org/10.1073/pnas.96.12.6745

[20] Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C. and Lander, E.S. (1999) Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science, **286**, 531-537.

[21] Shipp, M.A., Ross, K.N., Tamayo, P., Weng, A.P., Kutok, J.L., Aguiar, R.C.T., *et al*. (2002) Diffuse Large B-Cell Lymphoma Outcome Prediction by Gene-Expression Profiling and Supervised Machine Learning. *Nature Medicine*, **8**, 68-74. https://doi.org/10.1038/nm0102-68

[22] Yang, K., Cai, Z., Li, J. and Lin, G. (2006) A Stable Gene Selection in Microarray Data Analysis. *BMC Bioinformatics*, **7**, Article No. 228. https://doi.org/10.1186/1471-2105-7-228

[23] Huang, H., Wu, N., Liang, Y., Peng, X. and Shu, J. (2022) SLNL: A Novel Method for Gene Selection and Phenotype Classification. *International Journal of Intelligent Systems*, **37**, 6283-6304. https://doi.org/10.1002/int.22844